심리학석사학위논문

On the Robustness of Factor Analysis Models : A Comparison between Traditional and Bayesian Factor Analysis Models

요인분석 모형의 강건성에 대하여: 전통적 모형과 베이지안 모형의 비교

2013년 2월

서울대학교 대학원

심리학과 계량심리 전공

박 준 석

On the Robustness of Factor Analysis Models

A Comparison between Traditional and Bayesian Factor Analysis Models

요인분석 모형의 강건성에 대하여

전통적 모형과 베이지안 모형의 비교

지도교수 김 청 택

이 논문을 심리학석사 학위논문으로 제출함.
2012년 12월

서울대학교 대학원
심리학과 계량심리전공
박 준 석

박준석의 심리학석사 학위논문을 인준함.
2012년 12월

위 원 장 박 주 용 (인)
부위원장 고 성 룡 (인)
위 원 김 청 택 (인)

Abstract

In this study, model robustness was examined for mainly two factor analysis models, TFA(Traditional Factor Analysis) and BCFA(Bayesian Copula Factor Analysis). There were three abnormal data scenarios, which were outlier, kurtosis, and high correlation matrix cases. Both models were applied to each of the scenario data. It was revealed that BCFA model outperforms TFA model across the scenarios: the former was superior in terms of robustness when compared to the latter. In fact, BCFA could resolve the 'big loading' problems which arise when TFA is applied to the dataset, revealing the factor structure clearly. Additionally, some related issues are discussed in this article.

Keywords: factor analysis, robustness, Bayesian statistics, copula, mixed model, outlier, kurtosis, high-correlation matrix.

Student number: 2011-20123

Table of contents

LIST OF TABLES

LIST OF FIGURES

Introduction

Factor analysis has been one of the most prevalent multivariate statistical method in the field of psychology since Spearman (1904). It has been reported that the largest portion of the articles of the journal *Psychometrika* had been devoted to the study of factor analysis (Nunnally, 1978), and the number of studies using the technique had increased geometrically (Comrey, 1978). In fact, one can search more than 60,000 results about factor analysis, by the keyword ‘Factor analysis’ using SCOPUS. Clearly it is a powerful and handy tool to use, which is routinely used in practice.

Since a factor analysis is a statistical model with assumptions, the effect of violations of them must be an important issue to deal with. One desirable feature of statistical models is robustness. Robustness refers to the ability of statistical procedures to resist to some degree of violations of assumptions: departure from normality, outliers, extreme skewness and kurtosis, etc. However, despite the enduring popularity of factor analysis, there had been limited number of studies which directly investigate the problem of robustness, especially as to factor analysis. Some studies come from the works of Yuan et al. (1998, 2001, 2004). In these articles, the authors argue that SEM (structural equation model) is not robust, especially to outliers, skewed data, extreme kurtosis. Considering the fact that factor analysis model is a special case of SEM, it is natural to assume that factor analysis models are not robust to outliers, positive or negative skewed data, and abnormal kurtosis. Such considerations will be explicitly addressed in this article.

In this study, some present exploratory factor analysis models, including traditional models

and Bayesian ones which include recently developed one, will be evaluated in terms of robustness. Specifically, the following three concepts related to robustness will be covered in this study: outliers, skewness, kurtosis, high-correlation covariance (correlation) matrices. The former three factors are concerned with the normality assumptions, which is especially critical when maximum likelihood estimation method is used. And the last theme, sample covariance (correlation) matrix which shows high correlation between some variables, is thought to be concerned with ease of estimation and fit indices. It is reported that when the correlations among variables are extremely high, estimation of parameters is not so easy and model fit indices do not work well (Browne et al, 2002). In this study, the problems regarding such concepts will be covered. The following factor analytic models will be considered.

Factor analysis models

Traditional factor analysis model (TFA)

Factor analysis models aim to explain observed (‘measured’, or ‘manifest’) variables in terms of fewer latent variables, and to reveal the dependency structure among the variables. The latent variables are called ‘factors’. This idea is reflected in the model specification of factor analysis:

Let X be a p-dimensional observed column vector variable. In other words, there are p observed variables. The following equation represents the basic idea of factor analysis:

$$X = \Lambda f + \varepsilon \quad (1)$$

Where Λ is the factor loading matrix, f is p-dimensional factor score vector, and ε is the unique factor score. We need some assumptions to derive factor analysis model from the equation above:

1. $E(F)=0$, $E(\varepsilon)=0$.
2. $Cov(\varepsilon, F)=0$. (i.e. factor score and error terms are uncorrelated)
3. $Cov(\varepsilon) = D_\Psi$. (i.e. error terms are not correlated to each other.)

With these assumptions in mind, it can be shown mathematically, that the following equation holds:

$$(2)\ \Sigma = \Lambda\Phi\Lambda + D_\Psi$$

Where Σ is the covariance/correlation matrix, Λ is the factor loading matrix, and D_Ψ is a diagonal matrix which contains unique variance terms as its diagonal elements. Λ and D_Ψ can be estimated in numerous ways, among which maximum likelihood estimation, ordinal/generalized least square methods are the most frequently used methods in practice. Many statistical packages, including SPSS and Mplus, provide such procedures.

In this article, this factor analytic model and other relevant factor analytic models (for example, oblique factor models) will be called as 'traditional factor analysis model', abbreviated as 'TFA', for convenience.

Bayesian factor analysis model (BFA)

Many traditional (frequentist) statistical models have their own Bayesian counterparts. Similarly, TFAs have their own, too. There exist Bayesian methods of estimating Λ and D_Ψ. Factor analytic models adopting Bayesian methods are called 'Bayesian Factor Analysis' models. The first BFA model was formulated by Press (1972, 1982). The initial Bayesian factor model used Wishart Distribution as the likelihood function of covariance matrix and used vague prior. And he used numerical method to estimate model parameters. Since then, many methodological advances was made by Kaufman and Press(1973), Martin and McDonald(1975), Wong(1980), Lee(1981), Euverman and Vermulst(1983), Mayekawa(1985), Shigemasu(1986), Akaike(1987), Arminger and Muthen(1998), Shi and Lee(1998), and so on. Recently, Murray et al. (2012) adopted the copula modeling in their Bayesian copula factor model, which will be discussed in detail later in this article.

There have been much methodological advances and breakthroughs in the area of Bayesian statistics, owing to some breakthroughs of MCMC (Marcov Chain Monte Carlo) methods: MCMC sampling procedures are now routinely used in BFA models to estimate parameters. Like many other Bayesian statistical models, BFA models can benefit from the advances of MCMC sampling methods, too. A new MCMC procedure called PX Gibbs sampler is being used in many BFA models, and the BFA models investigated in this article are not the exception. They adopt PX Gibbs sampler, too.

But not much, if any, attention was given to BFA models in the field of psychology: Even the presence of BFAs is not mentioned frequently from anywhere in the field of psychology.

There was just one article regarding BFA model in the journal Psychometrika, and no article from Journal of mathematical psychology. Such interest is just budding (Muthen et al, 2012). Considering the popularity and prevalence of both Factor analytic models and Bayesian statistics in modern psychology and cognitive sciences, this is quite a strange situation. So one objective of this article is to draw attention to BFA models, thus enriching the tools psychologists can use.

In the present study, two BFA models will be assessed: Gaussian factor model and Copula factor model. The former is more similar to the original formulation of BFA, and the latter quite different from the original model, and semiparametric in nature. The latter is the main theme of the next section.

Bayesian Copula Factor Model (BCFA)

Recently, a BFA model which adopts 'copula' as the mean of estimating dependency structure was formulated by Murray et al(2012). A copula is defined mathematically as follows:

Definition. Suppose y_s are the observed variables. A p-dimensional copula C is a distribution function on $[0, 1]^p$ where each univariate marginal distribution is uniform on [0, 1]. Any joint distribution F can be completely specified by its marginal distributions and a copula; that is, there exists a copula C such that

$$F(y_1, \ldots, y_p) = \mathbb{C}(F_1(y_1), \ldots, F_p(y_p))$$

where F_j are the univariate marginal distributions of F.

The main advantage of using 'copula' in statistical analysis is that it can independently estimate the dependency structure and univariate margins, even when categorical variables are included in the analysis, i.e. the analysis deals with mixed data. The copula C encodes dependence structure independent from marginal distributions, so there is no concern about confounding between marginal distributions and dependency structure, which is very common problem of the models which do not use copulas. The estimation for the copula C is done using extended rank likelihood (Hoff, 2007), the extended version of marginal rank likelihood. Advantage of the use of copula is that it can handle discrete variables, in contrast with marginal rank likelihood, which cannot handle discrete margins. In short, Bayesian copula factor model is especially good when handling datasets which contain categorical (ordinal) variables.

The setting of prior distributions on the factor loadings and the procedure of posterior inference is as follows.

Prior distribution. Murray et al(2012) used GDP(Generalized Double Pareto) prior. The prior distribution has the density

$$\pi(\lambda_{jh}) = \frac{\alpha}{2\beta}\left(1 + \frac{|\lambda_{jh}|}{\beta}\right)^{-(\alpha+1)}$$

Which has two parameters α and β, i.e. $\lambda_{jh} \sim \mathrm{GDP}(\alpha, \beta)$. If we take $\alpha = 3$ and $\beta = 1$, then the mean becomes zero, variance being 1, and $\mathrm{P}(-2 < \lambda_{jh} < 2) \approx 0.96$, whose behavior mimics the standard normal distribution.

Posterior inference. In this model, PX (Parameter-extended) Gibbs samping procedure is used to estimate model parameters. The main characteristic of this procedure is addition of redundant parameters, which helps solving autocorrelation problems. According to Liu and Wu (1999) and Meng and Van Dyk(1999), PX-Gibbs sampler`s mixing behavior is at least as good as original Gibbs sampler, and sometimes outperforms it. Description of the entire inference procedure will be too tedious, so it`ll be omitted here. See Murray et al(2012) for further discussion.

Domains of comparison

Outliers. Outliers arise frequently in practice, due to mistyping, or unfaithful response, etc. Outliers can affect the normality of the data seriously, and even just one outlier can break the normality of the data if it is big enough. In case of multivariate normal distribution, just a big outlier can affect the correlation or covariance matrix. To illustrate how this works, see the following example:

Example. Consider the following correlation matrix (n=100) of 10 variables. Each variable is assumed to follow $N(0,1)$.

Table 1. *The correlation matrix*.

1
0.6, 1
0.6, 0.7, 1
0.7, 0.6, 0.8, 1
0.3, 0.3, 0.3, 0.2, 1
0.2, 0.1, 0.2, 0.1, 0.6, 1
0.2, 0.3, 0.3, 0.3, 0.7, 0.7, 1
0.2, 0.3, 0.3, 0.3, 0.3, 0.3, 0.4, 1
0.25, 0.4, 0.2, 0.25, 0.2, 0.1, 0.2, 0.6, 1
0.1, 0.1, 0.3, 0.2, 0.2, 0.15, 0.2, 0.7, 0.7, 1

The 10 variables above represent the correlation matrix of a multivariate dataset, which has 3-factor structure. From the first to fourth variables are bound to the first presupposed latent variable (factor), and the fifth to seventh to the second factor, and the rest to the third factor.

When an outlier is inserted to the dataset, what will happen? To see the effect, a multivariate normal dataset is generated, using the correlation matrix above, by using *mvrnorm* function in the MASS package of the statistical software R. And a single case of the first variable is replaced by an ‘outlier’, which was equal to 100 sd units. (In fact, this magnitude can be thought as too unrealistic, but to illustrate the effect of outliers more clearly, such a big outlier was chosen.) After this manipulation, the resulting correlation matrix was computed. The correlation matrix is given below (rounded off to two decimal points.)

Table 2. *The correlation matrix after introduction of outliers into the dataset.*

1
0.20, 1
0.09, 0.76, 1
0.14, 0.62, 0.78, 1
-0.01, 0.50, 0.42, 0.33, 1
0.08, 0.17, 0.17, 0.06, 0.58, 1
0.07, 0.41, 0.36, 0.38, 0.72, 0.66, 1
0.08, 0.32, 0.34, 0.40, 0.27, 0.16, 0.41, 1
0.10 , 0.37, 0.18, 0.18, 0.24, 0.10, 0.23, 0.65, 1
-0.08, 0.06, 0.23, 0.17, 0.11, 0.11, 0.20, 0.74, 0.71, 1

As can be seen from the matrix, the correlation between the first variable and the second variable fell down dramatically. The situation is similar, seeing the correlation between first and third variable. Considering the nature of TFA models, this pattern is critical to the analysis, since traditional FA models receive only correlation/covariance matrix as input. This is clearly a very problematic situation for factor analytic studies.

This problem will not cease to be harmful, even when the outliers are detected: One should decide how to treat the outliers. Simply eliminating the whole case is the worst thing to do, as noted by many statistics textbooks. Such treatment is especially bad when the number of data is not large, since the value of each data can be extremely high. In fact, it is also a difficult matter as to how to define ‘outliers’: there is no explicitly agreed criterion to use when defining ‘outliers’. Finally, the size of the data can be too large to detect outliers

efficiently. So even if we try to detect and get rid of outliers, it is not an easy problem.

But for BFA models, there is another possibility. If FA models do not use the covariance matrices as the input and use the raw dataset itself as the input, situation could be much different. This is the case of BFA models. So we can expect the BFA model to be more robust to outliers than TFA models. The issue will be addressed shortly.

Kurtosis. Kurtosis is formally defined as follows:

$$\beta_2 = \frac{\mathrm{E}(\mathrm{X}-\mu)^4}{(\mathrm{E}(\mathrm{X}-\mu)^2)^2} = \frac{\mu_4}{\sigma^4}$$

Where E is the expectation operator, μ is the mean, μ_4 is the fourth moment about the mean, and σ is the standard deviation. The normal distribution has a kurtosis of 3, so $(\beta_2 - 3)$ is used as indicator of positive/negative kurtosis. (DeCarlo, 1997) A normal distribution with large kurtosis has heavy tails and higher peak, and one with small kurtosis has light tails and relatively flat peak. A probability distribution with excessively high or low kurtosis crosses the curve of normal distribution with the same mean, twice.

There have been few studies which directly examine the effect of kurtosis on FA models. In this study, the problems related to kurtosis will be investigated. As will be seen shortly, departure from kurtosis 3 has detrimental effect on the quality of FA models.

High correlations among variables. A problematic situation is reported when the absolute values of some entries of correlation matrices are extremely large (Browne, 2002). In such

situations, parameter estimation frequently goes wrong, as will be seen. Abilities of FA models to recover the original structure under this condition will be discussed.

Model comparison

The three models discussed are evaluated in terms of robustness. Hypothetical datasets which have violations of assumptions of factor analysis were generated, and the three FA models were applied to the datasets. For TFA models, CEFA 3.04 package (Browne et al) is used, and for BFA(including copula model)s, bfa package(R package) is used. This package can be downloaded to R packages directly.

Study 1

In this study, model robustness against outliers is examined for three FA models. The procedure and results are given below.

Procedure. To examine model robustness, an imaginary factor structure was created in advance, and a multivariate normal dataset was generated from the implied correlation matrix which reflects the factor structure.. After the dataset was created, outliers were introduced to the dataset, and factor analysis models were applied to it. Model robustness was evaluated in terms of the ability of the models to restore the original factor structure under presence of the outliers.

The virtual factor structure is presented below.

Table 3. *Factor loadings of the virtual factor structure.*

Variable	F1	F2	F3
V1	0.7	0.2	0.2
V2	0.7	0.2	0.2
V3	0.7	0.2	0.2
V4	0.2	0.2	0.2
V5	0.2	0.7	0.2
V6	0.2	0.7	0.2
V7	0.2	0.7	0.7
V8	0.2	0.2	0.7
V9	0.2	0.2	0.7
V10	0.2	0.2	0.7

The correlation matrix. The correlation matrix was created from the implied covariance matrix, which had been calculated from the factor loadings above. The matrix is presented below.

Table 4. *Correlation matrix of the variables.*

1									
.55	1								
.55	.69	1							
.26	.33	.33	1						
.30	.38	.38	.57	1					
.30	.38	.38	.57	.66	1				
.27	.34	.34	.28	.33	.33	1			
.31	.39	.39	.33	.38	.38	.60	1		
.31	.39	.39	.33	.38	.38	.60	.69	1	
.35	.43	.44	.37	.43	.42	.67	.78	.77	1

Two virtual cases were considered in this study : when there is just one big outlier, and when there are many moderate outliers. Manipulation of the former case is exactly the same as one already discussed (table #2), and for the latter case, five outliers were inserted in the dataset whose magnitude was not as large as one used in the former case. Three factor analysis methods were then applied to the datasets. The results are presented below.

Single, large outlier case

TFA result. TFA was first applied to the dataset and orthogonal rotation (varimax) was applied later to get ‘clear’ factor structure. The following table shows the values of factor loadings after varimax rotation.

Table 5. *TFA factor loadings for single, big outlier dataset.*

Variable	F1	F2	F3
V1	.28	.08	.12
V2	.31	.10	.73
V3	.22	.10	.97
V4	.11	.64	.12
V5	.25	.89	.05
V6	.36	.71	.21
V7	.66	.16	.27
V8	.73	.23	.37
V9	.72	.27	.25
V10	.79	.38	.27

As can be seen from the table above, F3 no longer loads heavily on the first variable, which had to be bounded to the factor. This shows that just a big outlier is enough to distort the factor structure. (In addition, V1 is not being explained enough by any factor present.) In addition, a factor loading of V5 seems to be too high (.89), compared to the original factor structure. It should have been .7, but it is not. This problem, in which too big loadings emerge somewhere in the factor loading matrix, will be referred to ‘big loading problem’ from now on.

BFA result. Two BFA models were applied to the dataset, ordinary Gaussian factor model and

Gaussian copula factor model. The number of MCMC simulation prior to sampling was set to 1000, and 100 samples were discarded as burn-in. (It is interesting to note that too much simulation led to a very bad result.) After that, factor loadings were rotated using varimax rotation for ease of interpretation, by using CEFA package. (CEFA package provides a method of just rotating factor loadings.) The results are given in the following table 4. In fact, the Gaussian factor model did not work well, so the results will be excluded from further studies. From now on, the term 'BFA' will be used to refer to only BCFA model.

Table 6. *BFA factor loadings for single, big outlier dataset.*

Variable	F1	F2	F3
V1	.23	.63	.05
V2	.22	.83	.10
V3	.26	.79	.10
V4	.08	.10	.62
V5	.29	.03	.74
V6	.34	.19	.68
V7	.62	.29	.14
V8	.70	.34	.18
V9	.70	.34	.18
V10	.70	.20	.23

As can be seen on the table 4, the detrimental effect of the outlier is resolved. The intended factor structure was clearly revealed by BFA. V1 is not excluded from the variable group to

which the second factor loads heavily. One more advantage of this analysis is that extreme factor loadings do not emerge. In case of TFA, V5 was representative of big loading problem. It was peculiar to observe such result, considering the original factor structure. This problem is resolved in this analysis: the factor loading value is deceased to 0.74, roughly equal to the original value .7.

Multiple, moderate outliers case

In the present case, five moderate outliers were introduced to the dataset. In the preceding analysis, the magnitude of big outlier was 100 standard deviation units. In this case, the magnitudes of outliers were set to 10 standard deviations. After outlier generation, they were inserted into V1~V10, respectively. Then FA models were applied to the dataset. The results are given below.

TFA model. TFA was applied to the dataset. The factor loading matrices are presented.

Table 7.*TFA factor loading matrices for 10 outliers dataset.*

	Original structure			Outliers introduced		
Variable	F1	F2	F3	F1	F2	F3
V1	.7	.2	.2	.18	.63	.08
V2	.7	.2	.2	.23	.56	-.07
V3	.7	.2	.2	.23	.39	.06
V4	.2	.7	.2	.30	.05	.15
V5	.2	.7	.2	.10	.01	.99
V6	.2	.7	.2	.42	.10	.37
V7	.2	.2	.7	.39	.26	.18
V8	.2	.2	.7	.62	.18	.16
V9	.2	.2	.7	.66	.05	.06
V10	.2	.2	.7	.76	.22	.17

The factors 1, 2, 3 of the original analysis correspond to the factors 2, 3, 1 on the table, respectively. As we can see V6 on the table, due to the ten outliers inserted, the factor loadings were decreased to the extent that the variables are not explained by the factors enough, or even hard to determine to which the variable should be bound, compared to what is expected.

The communalities dropped as well, as can be seen on the tables below.

Table 8. *Communalities and unique variances of the preceding analysis.*

	Original analysis		Outliers introduced	
Variable	Communality	Unique variance	Communality	Unique variance
V1	.57	.43	.36	.64
V2	.57	.43	.43	.57
V3	.57	.43	.24	.76
V4	.57	.43	.12	.88
V5	.57	.43	.10	.90
V6	.57	.43	.19	.81
V7	.57	.43	.24	.76
V8	.57	.43	.45	.55
V9	.57	.43	.41	.59
V10	.57	.43	.67	.32

Table 4 shows how the communalities dropped from the original analysis. All the communalities declined simultaneously, indicating that the proportion observed variables are explained by the factors decreased. This indicates that this analysis was not robust to outliers.

When the magnitudes of deviation of outliers are stronger, the result of analysis becomes even worse. In the next analysis, the magnitude is set to 20 standard deviation units. The results are given below.

Table 9. *Comparison between 10 and 20 outliers: factor loadings*

	10 outliers			20 outliers		
Variable	F1	F2	F3	F1	F2	F3
V1	.18	.63	.08	.08	.38	-.01
V2	.23	.56	-.07	.26	.48	-.07
V3	.23	.39	.06	.15	.26	.00
V4	.30	.05	.15	.33	-.18	.07
V5	.10	.01	.99	.06	.00	1.00
V6	.42	.10	.37	.38	.07	.27
V7	.39	.26	.18	.35	.29	.15
V8	.62	.18	.16	.53	.06	.12
V9	.66	.05	.06	.52	.12	.13
V10	.76	.22	.17	.53	.28	.13

From the table, we can observe similar peculiar behaviors of the factor loadings, making it harder to identify the factors. The communalities dropped once again as well. The following table shows the extent of decrease. As the number of outliers increase, the quality of analysis becomes worse. Note that the communality of V1 dropped to zero.

Table 10. *Communalities and unique variances of 10 and 20 outliers data.*

	20 outliers		10 outliers	
Variable	Communality	Unique variance	Communality	Unique variance
V1	.00	1.0	.36	.64
V2	.19	.81	.43	.57
V3	.07	.93	.24	.76
V4	.06	.94	.12	.88
V5	.15	.85	.10	.90
V6	.18	.82	.19	.81
V7	.24	.76	.24	.76
V8	.28	.72	.45	.55
V9	.29	.71	.41	.59
V10	.33	.66	.67	.32

To sum up, in case of TFA, it could be said that outliers significantly distort the factor structure, given that they are many and big enough. These problems are thought to arise from the effect of outliers on correlation (covariance) matrices : as we saw previously in the introduction section, just a few extreme outliers can change the correlation/covariance matrices significantly. This will be discussed again in the discussion section. To conclude, TFA models do not seem to be robust to the outliers.

Comparison with BFA model. The same dataset used in the preceding analysis was used in this analysis, too. Additionally, Copula factor model is applied to the dataset. The result and comparison with TFA result is given below.

When the outliers had been included in the dataset, the differences of analysis result between the two models were revealed. The following table contains factor loadings of two analyses. (Both models assumed 3-factor structure.)

Table 11. *TFA and BFA factor loadings for 10 outliers dataset.*

	Original loadings			TFA			BFA		
Variable	F1	F2	F3	F1	F2	F3	F1	F2	F3
V1	.7	.2	.2	.18	.63	.08	.22	.05	.60
V2	.7	.2	.2	.23	.56	-.07	.25	.10	.79
V3	.7	.2	.2	.23	.39	.06	.27	.14	.69
V4	.2	.7	.2	.30	.05	.15	.13	.54	.12
V5	.2	.7	.2	.10	.01	.99	.27	.75	.03
V6	.2	.7	.2	.42	.10	.37	.35	.68	.22
V7	.2	.2	.7	.39	.26	.18	.59	.20	.28
V8	.2	.2	.7	.62	.18	.16	.66	.25	.33
V9	.2	.2	.7	.66	.05	.06	.68	.23	.19
V10	.2	.2	.7	.76	.22	.17	.74	.32	.28

Two important results can be drawn from this table. As we can see directly, the effects of outliers are attenuated in BFA outcome than TFA outcome. First, this difference is manifest across the loadings. This difference is especially obvious in case of V5. From TFA results, it is obvious that the analysis failed to recover the original factor loading. But BFA did not: it recovered roughly the same loading. And if we compare the results of TFA and BFA, it is

clear that BFA recovered the original factor structure better than TFA.

In the next analysis, the number of outliers was increased to 20. The result is given below.

Table 12. *TFA and BFA factor loadings for 20 outliers dataset.*

	Original loadings			TFA			BFA		
Variable	F1	F2	F3	F1	F2	F3	F1	F2	F3
V1	.7	.2	.2	.08	.38	-.01	.18	.55	-.01
V2	.7	.2	.2	.26	.48	-.07	.25	.76	.06
V3	.7	.2	.2	.15	.26	.00	.28	.61	.09
V4	.2	.7	.2	.33	-.18	.07	.12	.03	.52
V5	.2	.7	.2	.06	.00	1.00	.27	-.06	.70
V6	.2	.7	.2	.38	.07	.27	.28	.19	.69
V7	.2	.2	.7	.35	.29	.15	.61	.26	.17
V8	.2	.2	.7	.53	.06	.12	.62	.28	.24
V9	.2	.2	.7	.52	.12	.13	.68	.18	.23
V10	.2	.2	.7	.53	.28	.13	.71	.25	.28

The result of BFA recovers the original factor structure significantly better. In contrast to the result of TFA, which is severely damaged by the 20 outliers, BFA result seemed to have been damaged by the 20 outliers. Notably, V5 does not seem to suffer from outliers, when it comes to BFA: it restored the factor loading, which is exactly the same as one in the original factor structure. These differences would eventually lead to the difference of communalities,

which will be confirmed in the following investigation.

The communalities of two models are presented on the table below.

Table 13. C*ommunalities of TFA and BFA models (10 outliers).*

	TFA		BFA		Difference
Variable	Communality	Unique variance	Communality	Unique variance	(BFA-TFA)
V1	.44	.56	.40	.60	-.04
V2	.37	.63	.69	.31	.32
V3	.21	.79	.57	.43	.36
V4	.12	.88	.33	.67	.21
V5	1.00	.00	.63	.37	-.37
V6	.33	.67	.62	.38	.29
V7	.26	.74	.46	.54	.20
V8	.45	.55	.61	.39	.16
V9	.45	.55	.55	.45	.10
V10	.66	.34	.73	.27	.07

The differences between communalities were all positive (mean difference = 0.186). Besides, the difference was not trivial. In some cases, the BFA communality was more than twice the corresponding TFA loading (V3, V4). To test the significance of differences, An one-sample t-test was conducted. (V5 was excluded from the analysis, due to the presence of factor loading 1.) The mean difference was significantly different from zero at α=0.05(two-sided), t(8)=4.3267, p=0.002523, r^2=0.70. In the present analysis, BFA model outperformed

the TFA model.

When it comes to the dataset which has more outliers (twenty), the difference becomes even greater, seeing the table below.

Table 14. *The communalities of TFA and BFA models (20 outliers).*

	TFA		BFA		Difference
Variables	Communality	Unique variance	Communality	Unique variance	(TFA-BFA)
V1	.15	.85	.33	.67	.18
V2	.30	.70	.64	.36	.34
V3	.10	.90	.46	.54	.36
V4	.15	.85	.29	.71	.14
V5	1.0	.00	.56	.44	-.44
V6	.22	.78	.60	.40	.38
V7	.23	.77	.46	.54	.23
V8	.30	.70	.52	.48	.22
V9	.30	.70	.55	.45	.25
V10	.37	.63	.65	.35	.28

Mean difference between TFA communality and BFA communality was 0.264, which is bigger than 0.186 of the previous result. To test the significance of this difference, An one-sample t-test is conducted once again. The difference was significantly different from zero at α=0.05(two-sided), t(8)=9.6186, p<0.001, r^2=0.92. Seeing the effect size r^2, it could be said that the more outliers, the more TFA model departs from the original factor structure

than BFA model.

To summarize, as the number and intensity of outliers gets bigger, the quality of TFA analysis becomes worse, as we could see from the analyses. But this was not the case of BFA models, suggesting that BFA model is more robust to the outliers.

Study 2

In this study, the problems related to kurtosis are discussed. As noted, a distribution with small/large kurtosis is characterized informally as whose density crossing that of the standard normal distribution twice. This behavior affects the thickness of the tails of distributions, and sharpness of the peak, and threatening normality of the data. So it is desirable for the FA models to resist such problems. Robustness to kurtosis is tested across the models, and compared.

Procedure and results

High kurtosis case. A dataset with large kurtosis is generated through *mvrnorm* function of the R package, generating multivariate dataset repeatedly while the desired dataset with at least one variable with small/high kurtosis is acquired. By this procedure, a dataset with two variables whose kurtosis is high was generated. The correlation matrix is given below. The factor structure of the data was identical to the preceding analyses, in which variable 1~4 are bound to the factor 1, and 5~7 to the factor 2, and 8~10 to the factor 3.

Table 15. *The correlation matrix of high kurtosis case.*

1									
0.55	1								
0.56	0.72	1							
0.64	0.56	0.80	1						
0.33	0.44	0.32	0.18	1					
0.27	0.22	0.31	0.23	0.65	1				
0.18	0.30	0.35	0.40	0.68	0.74	1			
0.27	0.33	0.35	0.31	0.45	0.37	0.44	1		
0.24	0.41	0.19	0.17	0.28	0.12	0.20	0.62	1	
0.10	0.15	0.30	0.15	0.30	0.23	0.27	0.69	0.74	1

The kurtosis of each variable was 3.17, 5.62, 3.87, 3.04, 4.21, 2.86, 2.78, 3.04, 2.44, 3.05, respectively. Departure from kurtosis 3 indicates the degree of abnormality of the variable, and significance of these departures can be tested statistically, by the procedure 'Anscombe-Glynn test of kurtosis'. The null hypothesis of this test is that the kurtosis of data is equal to 3, so if the null hypothesis is rejected, then we can regard the data as non-normal. When this test was applied to the 10 variables, two of them were found to have non-normal kurtosis. They were second and fifth variable, p=.001, .032, respectively. To show the pattern of the high-kurtosis data, two histograms of V2 and V5 are given below.

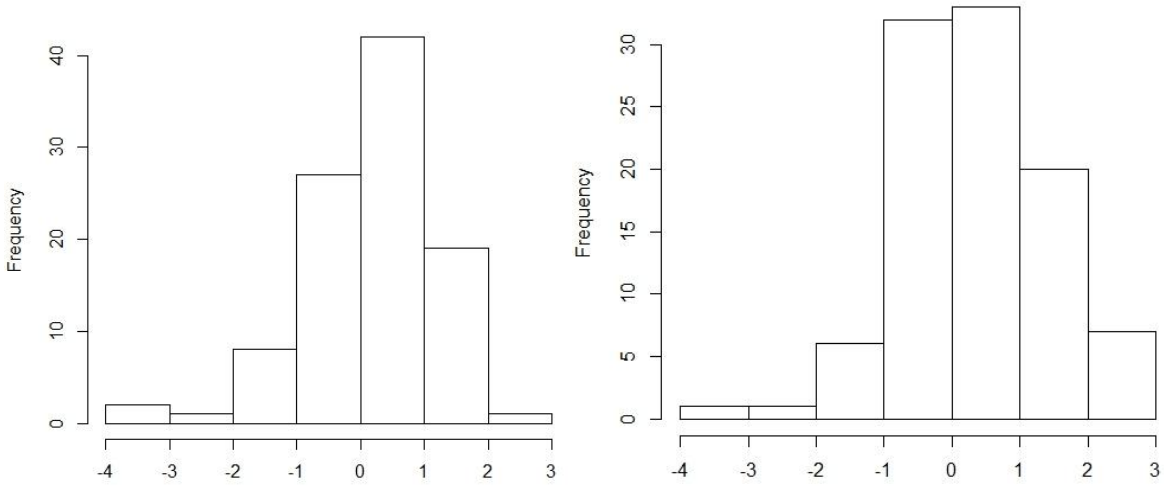

Figure 1. *The histogram of V2 and V5*

The two factor analysis models were then applied to the datasets. The results are given below.

TFA result. Interestingly, CEFA failed to give the result when iteration limits was set to 50. So the iteration limits was increased by 50, once a time. CEFA gave first result when it reached 200. In other words, it did not give result when the iteration limit was 150. The result is in the following table.

Table 16. *TFA result with 2 variables with high kurtosis.*

Variables	F1	F2	F3
V1	.61	.14	.17
V2	.70	.14	.34
V3	.93	.19	.08
V4	.82	.16	.08
V5	.18	.73	.22
V6	.15	.84	.06
V7	.21	.83	.13
V8	.23	.38	.58
V9	.10	.06	.99
V10	.14	.20	.71

Considering the correlation matrix, it is strange to observe extremely high loadings, which are above 0.9. And the phenomenon was manifest in the ninth variable : the third factor` s loading on the ninth variable is 0.99, which is not reasonable, considering the correlation structure. As we could realize from study 1, the presence of ‘big loading’ implies some failure of the model. This awkward conclusion is not expected.

BFA results. The following table shows the result of BFA to the same dataset.

Table 17. *BFA result with 2 variables with high kurtosis.*

Variables	F1	F2	F3
V1	.61	.13	.14
V2	.63	.27	.17
V3	.81	.13	.14
V4	.80	.06	.11
V5	.13	.21	.66
V6	.12	.09	.72
V7	.20	.14	.72
V8	.21	.65	.31
V9	.11	.79	.06
V10	.12	.73	.13

When compared to the result of TFA, this analysis gives more reasonable estimates. Implied factor structure from the table is clear, as expected. But the loadings no longer show extreme behavior: there is not abnormally big loading, compared to other loadings. Especially, the biggest loading of the V9 is that of the second factor, whose value is 0.79. This value is much attenuated, when compared to 0.99 of TFA analysis, demonstrating the robustness of BFA model clearly. And this is the case of V3, too. The greatest loading of V3 is 0.81, and this value seems to be more reasonable, when compared to the loading value 0.93 of the TFA result. These results clearly show that BFA model is more robust to the abnormal kurtosis than TFA model.

Low kurtosis case. The same analysis is conducted to low kurtosis dataset, generated by similar procedure used in creating high-kurtosis dataset. The correlation matrix is given below.

Table 18. *The correlation matrix of low kurtosis matrix.*

1									
0.56	1								
0.59	0.72	1							
0.69	0.65	0.83	1						
0.45	0.34	0.32	0.34	1					
0.36	0.02	0.16	0.23	0.63	1				
0.32	0.28	0.27	0.45	0.71	0.60	1			
0.30	0.17	0.22	0.34	0.26	0.23	0.31	1		
0.07	0.21	0.05	0.12	0.04	0.00	0.04	0.57	1	
0.06	-0.05	0.24	0.17	0.09	0.08	0.03	0.70	0.61	1

The kurtoses of the variables were 2.72, 2.65, 2.80, 2.93, 2.59, 3.29, 3.40, 2.65, 3.08, 2.06, respectively. And Anscombe-Glynn test revealed that the tenth variable`s kurtosis was too small to be normal (p=0.001084). The following figure shows the histogram of 10^{th} variable.

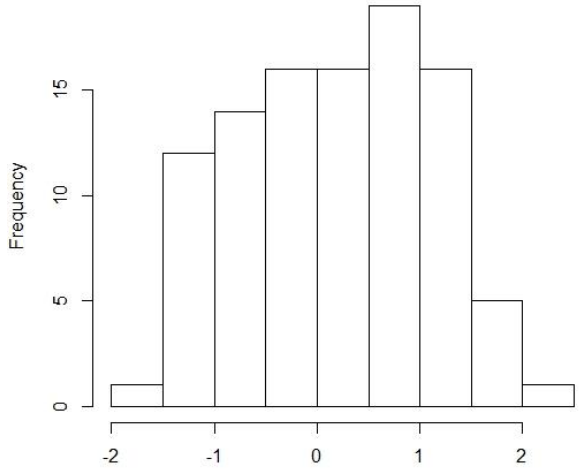

Figure 2. *The histogram of the tenth variable*.

Two FAs were conducted. The results are given below.

TFA result. The same problem which occurred in the preceding analysis occurred again. With just 50 iterations, CEFA failed to give solutions. When it was increased to 150, results were given for the first time. The result is given below.

Table 19. *TFA result with 1 variable with low kurtosis.*

Variables	F1	F2	F3
V1	.65	.33	.02
V2	.78	.11	-.10
V3	.91	.08	.18
V4	.87	.24	.12
V5	.24	.78	.07
V6	.06	.74	.08
V7	.24	.81	.02
V8	.17	.28	.69
V9	.04	.01	.61
V10	.07	-.01	1.00

The same problem can be perceived directly. Looking at the loadings of V10, we can know that the third factor loads to the variable too heavily. This is a strange result, considering the correlation structure. (There was no extremely high correlation.) The loadings of V3 seem to bear some problem, too. The first factor loads on V3 too heavily, though the magnitude is not as strong as the case of V10. In short, the same ‘big loading’ problems were observed, either the kurtosis was high or low.

BFA result. BFA is conducted, using the same dataset. The results are contained in the following table.

Table 20. B*FA result with 1 variable with low kurtosis.*

Variables	F1	F2	F3
V1	.07	.45	.17
V2	.14	.50	.05
V3	.26	.52	.15
V4	.20	.54	.11
V5	.37	.16	.53
V6	.25	-.02	.53
V7	.34	.18	.60
V8	.70	.15	.19
V9	.67	.14	.21
V10	.76	.09	.18

Clearly, the problem of ‘big loading’ is resolved. Seeing the loadings of V10, we can confirm that there is no extremely large factor loading. Although the overall absolute values factor loadings declined, the factor structure can be perceived clearly, indicating model robustness.

To sum up, it seems that, in both high and low kurtosis situations, BFA model is more robust than TFA model, resolving the ‘big loading’ problem.

Study 3

In this study, the problem arising from high correlations between variables is examined. Browne et al. (2002) addressed this problem, in the context of assessing the model fit. They found that even when the model fits to the data quite well, sometimes fit indices can be incompatible with this fact. According to Browne et al, this phenomenon is especially obvious when the unique variances of the observed variables are extremely small ; that is, communality of the observed variables are extremely high, suggesting very high correlation coefficients between the variables. The correlation matrix they used to illustrate the situation is given below (there were 8 observed variables) :

Table 21. *The correlation matrix from Browne et al. (2002)*

1
.902 1
.756 .862 1
.772 .891 .930 1
.114 .125 .147 .123 1
.095 .099 .114 .094 .959 1
.103 .111 .132 .115 .933 .988 1
.105 .104 .108 .092 .910 .981 .987 1

As we can figure out from the matrix above, there seem to be two factors: the first factor seems to load heavily on the variables #1~#4, and the second factor loads heavily on the variable #5~#8. The correlation coefficients are usually large, and especially among the

variables which are bound to the second factor: Their absolute values are all greater than 0.9, and two of them reached almost about 1(0.987, 0.988). In this situation, residual correlation coefficients are very small, and a problem arises in this case: fit indices indicate that the model fits very poorly to the data. This awkward conclusion is not expected for robust FA procedures, of course. Furthermore, estimation is sometimes not easy in this situation, like in the preceding study.

To examine robustness of the FA models, a dataset was created using the correlation matrix above. Two FA models are applied to this dataset. The results are given below.

Table 22. *The correlation matrix used in the analysis*

1								
0.91	1							
0.83	0.96	1						
0.24	0.34	0.26	1					
0.13	0.18	0.15	0.69	1				
0.33	0.37	0.26	0.79	0.51	1			
0.00	0.10	0.11	0.09	0.18	0.08	1		
0.04	-0.03	0.05	0.08	0.01	-0.06	0.64	1	
-0.02	0.04	0.03	-0.02	-0.08	0.14	0.76	0.47	1

There are 9 variables and 3 factors. The correlation coefficients among the variables which are bound to the first factor (first to third variables) are quite large, and one of them is 0.96, a value whose absolute value is about to 1. Two FA models are applied to this dataset. The

results are given below.

TFA result. Similar problems which were uncovered during the previous analyses had been found. The following table summarizes the result.

Table 23. *TFA result with high correlation matrix.*

Variables	F1	F2	F3
V1	.91	.09	-.05
V2	.98	.18	.04
V3	.96	.10	.06
V4	.16	.99	.02
V5	.05	.69	.13
V6	.24	.76	.02
V7	.05	.06	1.00
V8	-.07	.08	.64
V9	.02	-.04	.76

As can be seen from the table above, there are some loadings which are quite strange. Look at the loading of second factor on v4: it equals to 0.99, which is very high. This is not the sole case: The same phenomenon occurs at the factor loading of third factor on v7: it equals to 1.00. Even if we consider that it`s just a rounded-off value, this value implies some problem is present. Even though we appreciate the fact that high loadings of first factor on v1~v3 as arising naturally, since correlations among them were originally large, these abnormal

behaviors of factor loadings question the robustness of TFA model.

BFA result. The loadings of BFA analysis is summarized on the table below.

Table 24. *BFA result with high correlation matrix.*

Variables	F1	F2	F3
V1	.83	-.03	.09
V2	.91	.05	.18
V3	.89	.07	.10
V4	.20	.04	.79
V5	.08	.05	.61
V6	.25	.07	.67
V7	.07	.79	.09
V8	-.01	.57	.02
V9	.05	.70	.00

The behaviors of loadings indicate clear separation of the variables into three factors, but the ‘big loading’ problem does not seem to occur, unlike the case of TFA result. The biggest loading on the table is 0.91, but this is surely anticipated phenomenon, due to high correlations among V1~V3. But more important is the fact that other loadings are not affected by high correlations among V1~V3. This result suggests that BFA model is more robust to ‘high correlation data, than TFA model.

General discussion

Summing up the results from the studies, it seems that Bayesian Copula Factor analysis (BFA) is robust to the departures from normality of the data, or unusual high correlations between the variables. There was no sign of abnormality, especially ‘big loading’ problems, in the analysis results of BFA, at the same time recovering the intended factor structure efficiently.

What is the source of this robustness of BFA? This important question does not seem to be able to be answered entirely from this study. But a preliminary conjecture could be made. The answer could come from the inputs which the two models require: they are different from each other. The sole required information (raw data) when doing TFA is correlation/covariance matrix, which does not consider any other information, like the individual characteristics of the data and loadings. So when this correlation/covariance matrix is damaged severely, the result goes wrong, irreversibly: this phenomenon was especially manifest in case of outlier study. Just a single outlier was enough to distort the entire correlation matrix, and this led to the awkward result.

One peculiar phenomenon which was manifest across the studies was ‘big loading’ problem. The problem was observed across situations. It is not clear why, and how these big loadings are created, but it seems that the occurrence of big loadings may be a significant indicator of bad analysis, since there is no clear reason as to why the variable deserves such a big loading from just a single factor, at the same time the other factors virtually do not load on them. And other variables which are bound to the same factor could bear such big loadings.

Why ‘the’ variable should be loaded by a specific factor? This is not clear.

In contrast, such problems did not arise when it comes to BFA model. BFA model revealed the underlying factor structure clearly, at the same time it was free from ‘big loading’ problem, when the data were not normal. One thing to consider when finding the source of robustness is the nature of BFA model: it uses full dataset as the input, and puts prior distributions on factor loadings. Even though the model assumes the same model as TFA ($\Sigma = \Lambda\Phi\Lambda + D_{\Psi}$), these factors may reduce the harmful effect of outliers, high/low kurtosis, etc. One important feature of the bayesian inference is that, by applying Bayes`rule, prior information and information which comes from the data are reconciled. In many textbooks, posterior distributions are depicted as ‘the compromise between prior and data’, and maybe this proposition could be applied to BFA model. It may the case that because of many other ‘non-outliers’ presence, the data contained in outliers were overridden. Since this conjecture is merely a guess, the questions regarding the source of BFA’s robustness seem to need more investigation, in depth.

There is one more issue about BFA. It enables the users to construct mixed models. Mixed FA model is one which can incorporate both continuous and categorical variables. In this case, ‘categorical variable’ means ordinal variables, since nominal variables do not have ‘magnitude’. The problem is that some ordinal scales are treated as continuous in practice routinely, without explicit reason. But there have been some arguments which suggest that ordinal scales should not be treated as continuous variables. Some of such literatures are Knapp(1990), Kuzon et al(1996), Jamieson(2004). According to these authors, likert scale is simply not continuous scales, so should not be treated as interval or ratio scales. In these

arguments are valid, it`s so peculiar to use them to do FA studies.

It is not easy to implement mixed factor analysis using traditional methods. A FA analysis which uses some ordinal variables as input clearly violates the model assumption: there can be no expected value of ordinal variables. So one should find the route around, and such procedures are implemented in many routinely used statistical packages, like Mplus. (it is interesting to note that one of the most popular SEM packages, AMOS, does not provide solutions to such situations.) But it is not easy to use this program, and requires more effect to learn and use it.

By using BFA procedure, this problem can be resolved. BFA provides a mixed analysis procedure, and by just telling it which variables should be regarded as ordinal, we can conduct mixed factor analysis. Specifically, it solves the problems arising from using both continuous and ordinal data by adopting extended rank likelihood, which is an approximation to full likelihood. Using this likelihood for the estimation of copula, we can handle both discrete and continuous marginal distributions, and this is done easily with existing R package, bfa.

First significance of this study is that it directly addressed the robustness of existing factor analysis models. As noted, such studies are scarce: only limited numbers of studies tackle the problem directly. But this is not a trivial problem, since such non-normal data or highly correlated data do arise in practice, as we have seen from the study by Browne et al(2002). Espacially, many textbooks describe kurtosis as a kind of descriptive statistic, but in many cases it is simply ignored or not mentioned in practice, but it can have significant effect on

the analysis, as we have seen. Maybe more attention should be given to such considerations.

Second significance of this study is that it compared traditional factor analysis model and Bayesian factor analysis model. It seems that interest to Bayesian statistical methods are increasing, and it is time to learn more about the strength and weaknesses of both traditional (frequentist) and Bayesian methods. As reflecting the need to explore more about the possibilities of Bayesian statistics, an article was published in the *Psychological methods* recently (Muthen et al, 2012), which explores the possibility of incorporating Bayesian inference into one of the most used procedure, Structural Equation Modeling. Determining which model to use is not easy at present, and studies concerning such problems seem to need some investigation.

Conclusion

In the present study, robustness of TFA and BFA(especially Bayesian copula factor analysis) to some problematic situations have been examined. As we saw across the studies, BFA seems to be more robust to many abnormal situations than TFA. TFA revealed some problems when it was applied to such data, and the 'big loading' problem was the most manifest among the problems. But BFA did not show such weakness: it was robust to many problematic situations. So it could be said that BFA was more robust than TFA in this study.

Bayesian statistics is becoming popular in the field of psychology, not only just as methodological tools, but also as a research paradigm. (Griffiths & Tenenbaum, 2005, etc) Accurate depiction and assessment of Bayesian statistical methods are needed at this point of

time. This study is a preliminary try to such ambitious goal, but there remain many things to do, and they should be addressed gradually, as many other scientific programs did.

References

Browne, M. W., MacCallum, R. C., Kim, Cheong-Tag., Andersen, B. L., Glaser, R. (2002). When Fit Indices and Residuals Are Incompatible. *Psychological Methods*. 7(4), 403–421.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology*, 46(4), 648-659

DeCarlo, L. T. (1997). On the Meaning and Use of Kurtosis. *Psychological Methods*, 2(3), 292-307.

Jamieson, S. (2004). Likert scales: how to (ab)use them, *Medical Education.* 38: 1212–1218.

Johnson, R. A. & Wichern, D. W. (2002), Applied multivariate statistical analysis : 5th edition, *Pearson education International, Prentice Hall: Upper Saddle River.*

Knapp, T. R. (1990). Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy, *Nursing Research*, 39(2), 121-123.

Kuzon, W. M., Urbanchek, M. G., & McCabe, S. (1996), The seven deadly sins of statistical analysis, *Annals of Plastic Surgery*, 37(3), 265-272.

Murray, J. S., Dunson, D. B., Carin, L., & Lucas, J. E. (2012). Bayesian Gaussian copula factor models for mixed data, *arxiv.org*, working paper.

Press, S. J. (2003). Subjective and Objective Bayesian statistics : 2nd edition. *Wiley-Interscience : Hoboken, New Jersey.*

Spearman, C. (1904). General intelligence, objectively determired and measured. *American Journal of Psychology*, 15, 201-293

Yuan, Ke-Hai & Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical & Statistical Psychology,* 51 : 63-88.

Yuan, Ke-Hai & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis, *British Journal of Mathematical & Statistical Psychology*, 54, 161-175.

Yuan et al(2004), Structural equation modeling with heavy tailed distributions, *Psychometrika*, 69(3), 421-436.

국문 초록

본 연구에서는 전통적 요인분석 모형(Traditional Factor Analysis: TFA)과 베이지안 Copula 요인분석 모형(Bayesian Copula Factor Analysis: BCFA)의 강건성(robustness)에 대해 연구하였다. 세 종류의 모형 가정 위배 상황이 가정되었다: 그것들은 이상점의 존재(outliers), 비정상적인 첨도(kurtosis), 그리고 고상관 상관행렬(high correlation matrix) 였다. 두 요인분석 모형은 각각의 상황에 적용되었다. 연구 결과 세 경우 모두 BCFA가 강건성의 측면에서 TFA보다 뛰어난 수행을 보여 주었다. 특히 BCFA 모형은 데이터의 요인구조를 잘 추출하면서도, 동시에 TFA 모형으로는 해결하기 어려운 '큰 요인부하량 문제'(big loading problem)를 잘 해결할 수 있었다. 부가적으로 이와 관련된 몇 가지 사항들이 논의되었다.

주요어 : 요인분석, 강건성, 베이지안통계, copula, 혼합모형, 이상점, 첨도, 고상관 상관행렬

학번 : 2011-20123